Data Science Capstone Lab

Data Science 400 • Spring 2021

Time & Location:
Tuesdays & Thursdays 1:30pm - 3:00pm
Location: Fully Online - Synchronous (please check Canvas for link)

Instructor:
Arko Barman (ab139@rice.edu)

Office Hours:
Wednesdays 2:30 - 3:30pm or by appointment
Location: Fully Online (please check Canvas for link)

D2K Lab Administrator:
Jennifer Sanders (Jennifer.Sanders@rice.edu or d2k@rice.edu)

Course Description:
DSCI 400 is a capstone course where student teams will complete semester-long data science projects involving research and application using data of their choice (subject to approval from the instructor). Students will also learn best practices in data science.

Course Overview:
This capstone course is designed to provide students a hands-on experience in designing an end-to-end data science pipeline by working collaboratively in teams. The project will involve research of relevant literature and the application or development of data science methods/algorithms to solve a data science problem using a selected dataset. Each student team will select a dataset of their choice with the instructor's approval, formulate a data science problem, and work together to develop algorithms and code to solve it. The members of each student team are required to meet weekly among themselves outside of the class hours to discuss progress, delegate and divide up their tasks, and set both short term and long term goals for their project. During the semester, teams will work towards project milestones and present their progress in interim reports, oral presentations, and software checks. At the end of the semester, each team will deliver a final project report along with their software and give a presentation on their work.

This course also consists of a lecture component in the form of relevant lectures on various concepts throughout the semester, where the students will learn how to design a data science pipeline and relevant theoretical concepts in statistics, machine learning, version control systems, cloud computing, best practices in software development, and reproducibility in data science.

Course Objectives & Outcomes:
This course will provide an experiential learning environment for students, where they become familiar with the best practices in data science through hands-on experience by completing a team-based real-world data science problem. Students will master techniques and methods in data wrangling, exploratory data analysis, and modelling through the design and development of a reproducible data science pipeline; validate their findings, and communicate their results to a broad audience. By completing this project, students will also develop skills for working effectively in teams, leadership, project management, and different forms of communication. Students completing this course will be able to:

(i) successfully able to design and develop an end-to-end reproducible data science pipeline

(ii) successfully complete a data science project involving research, analysis, and application of data science and machine learning tools

(iii) work collaboratively in a team to complete a data science project

(iv) communicate their data science findings to both lay and technical audiences through written, oral, and visual communication

Course Application & Team Formations:
This course requires an application, where the students are required to elaborate on their data science experience through a series of questions. These questions are designed to enable the instructor to gauge the background and relevant experience of a student vis-à-vis other students and to form the most effective teams by leveraging the "jigsaw" cooperative learning strategy.

Lectures:
Every class will consist of a recorded lecture component of 30-45 minutes on a relevant topic in data science. The lectures in the initial classes will help students grasp the fundamental tenets of how to design a data science pipeline, how to effectively collaborate on developing a code base, best practices in software development, cloud computing, reproducibility in data science, exploratory data analysis, and other general data science topics. The topics for subsequent lectures are flexible and will be decided based on the needs of the students enrolled in the course and the projects undertaken by the student teams.

Weekly Team Updates:
After the teams are finalized and the datasets for the projects are decided, every team is required to provide a 2-5 minute weekly update on their project in class. These updates should include at least two team members in the presentation and additionally, over the course of the entire semester, each student must have participated in at least two weekly updates. The updates will not only help the members of every team in developing effective communication skills, but will also help the entire class in understanding the wide variety of data science problems that are being tackled by different teams in the class.

Help Session:
Subsequent to lectures and weekly team updates, the remaining class time will be devoted to team meetings with the instructor to discuss the progress, chart a path for the future, and other relevant discussions about technical aspects of the project. Every team will meet with the instructor at least once a week during a help session for discussions.

Coursework:
The major work of this course will be to complete a data science project, involving research, analysis of data, application, and communication. Each team will present their work through presentations at the end of the semester and a final report accompanied by software due at the end of the semester. Throughout the course of the semester, teams will submit interim reports, present their work in the class, and have software checks with the instructor. These interim reports, presentations, and software checks have been designed to provide feedback and guidance to the teams on the best practices to follow when submitting their final report, presentation, and software. Detailed instructions, rubrics, and checklist for presentation, report, and software are provided in a separate document.

Course Expectations:
Students are expected to work at least 10 hours per week on their project outside of class hours. All teams are required to have at least one meeting among their members every week to keep track of their progress, make short-term and long-term plans, and delegate work among themselves. Students may conduct these meetings over Zoom or in person; for in-person meetings, please follow the university guidelines and contact the instructor if you need any assistance.

Attendance at team meetings and classes is required unless you have made prior arrangements with the instructor. Moreover, each member of a team is expected to contribute to all aspects of the data science project, including but not limited to technical contributions, literature review, coding, and preparation of the report and presentations, and presenting in the class.

Grading Overview:
Individual: 30%
- Peer evaluations: 10%
- Individual project contributions: 10%
- Class Participation: 10%

Team: 70%
- Interim Evaluations: 15%

Final Report: 25%
Final Presentation: 15%
Final Software: 15%

Grading rubrics, checklists, and other useful details on all coursework are given in a separate document. Note that students who do not contribute in any meaningful way to the team project will receive a failing grade for this course regardless of the team grade.

Accommodations for students with disabilities:
Students with a documented disability requiring accommodations should speak with the instructors during the first two weeks of class.

This syllabus is subject to change with reasonable advance notice by the instructor.

Course Schedule:

Date	Class Time	Coursework Due
26-Jan	Intro + Course Logistics	
28-Jan	Lecture: Data Science Pipeline	
2-Feb	Meet your teammates! + Lecture: Designing a DS Pipeline + Hunt for Datasets	
4-Feb	Lecture: Data Science Reproducibility & Reliability	
9-Feb	Lecture: Best Practices in Software Development	
11-Feb	Lecture: Version Control System & Git	Dataset to be finalized (with approval from instructor)
16-Feb	Weekly Updates + Lecture: Cloud computing and AWS	
18-Feb	Weekly Updates + Lecture: Project Management	
23-Feb	Initial Presentations + Lecture: Exploratory Data Analysis	
25-Feb	Initial Presentations + Lecture: Exploratory Data Analysis	Initial Report
2-Mar	Weekly Updates + Lecture	
4-Mar	Weekly Updates + Lecture	
9-Mar	Software Check I + Lecture	
11-Mar	Software Check I + Lecture	OPTIONAL: Revised initial report due
16-Mar	NO CLASS: Sprinkle day	
18-Mar	Weekly Updates + Lecture	
23-Mar	Weekly Updates + Lecture	
25-Mar	Weekly Updates + Lecture	
30-Mar	Midterm Presentation + Lecture	
1-Apr	Midterm Presentation + Lecture	Midterm Report
6-Apr	Weekly Updates + Lecture	
8-Apr	NO CLASS: Sprinkle day	
13-Apr	Software Check II + Lecture	
15-Apr	Software Check II + Lecture	OPTIONAL: Revised midterm report due
20-Apr	Weekly Updates + Lecture	
22-Apr	Weekly Updates + Lecture	
27-Apr	Final Presentation Rehearsals	
29-Apr	Final Presentations	
6-May	NO CLASS	Final Report + Final Software Due